L16: Entry 4 of 6 File: USPT Jul 20, 1999

DOCUMENT-IDENTIFIER: US 5926811 A
TITLE: Statistical thesaurus, method of forming same, and use thereof in query expansion in automated text searching

Abstract Text (1):
A statistical thesaurus is built dynamically, from the same text collection that is being searched, allowing improved generation of expanded query terms. The thesaurus is dynamic in that thesaurus records are collected, ranked, accessed, and applied dynamically. Thesaurus "records" are actually formed as indexed documents arranged in "collections". The collections are preferably distinguished based on text source (court cases versus news wires versus patents, and so forth). Each record has terms assembled in indexed groups (or segments) which inherently reflect a ranking based on relevance to an initial query. After an initial query is received, the appropriate collection(s) of records may be searched by a conventional search and retrieval engine, the searches inherently returning records ranked by degree of relevance due to the record indexing scheme. A record ranking scheme avoids contamination of relevant records by less relevant records. The record selection and the expansion query term generation processes are each divided into parallel threads. The separate threads correspond to respective text sources to enable the improved expansion query term generation to be provided in real time.

Brief Summary Text (3):
The present invention relates generally to the field of automated search and retrieval of text documents. More specifically, the invention relates to thesauri (especially statistical thesauri), to the structures of the statistical thesauri, to methods of forming the statistical thesauri, and to use of the statistical thesauri in query expansion.

Brief Summary Text (5):
It is known in the field of information retrieval that both precision and recall can be greatly improved when queries are expanded to contain a larger number of good search terms. A thesaurus can be used to increase the number of good search terms.

Brief Summary Text (6):
A statistical thesaurus is a thesaurus which contains terms that are related to the headword by their co-occurrence with the headword in text. This is in contrast to a traditional thesaurus whose terms, synonyms, are related to the headword by meaning.

Brief Summary Text (7):
Recent research has shown that a statistical thesaurus provides good search terms when used for query expansion, while traditional thesauri provide little improvement and may actually hurt overall performance. As an example, FIG. 6 illustrates synonyms for the headword "murder" from a traditional thesaurus, while FIG. 7 illustrates the related concepts from a statistical thesaurus.

Brief Summary Text (11):
The inventive statistical thesaurus provides a high degree of performance, is scalable to multiple users and large amounts of source information, and is tunable to specific source information. The thesaurus works best when it is built from the text collection being searched. In order to meet these requirements the inventive dynamic, parallel thesaurus is provided.

Brief Summary Text (12):
A statistical thesaurus is built dynamically, from the same text collection that is being searched, allowing improved generation of expanded query terms. The thesaurus is dynamic in that thesaurus records are collected, ranked, accessed, and applied

dynamically. Thesaurus "records" are actually formed as indexed documents arranged in "collections". The collections are preferably distinguished based on text source (court cases versus news wires versus patents, and so forth). Each record has terms assembled in indexed groups (or segments) which inherently reflect a ranking based on relevance to an initial query. After an initial query is received, the appropriate collection(s) of records may be searched by a conventional search and retrieval engine, the searches inherently returning records ranked by degree of relevance due to the record indexing scheme. A record ranking scheme avoids contamination of relevant records by less relevant records. The record selection and the expansion query term generation processes are each divided into parallel threads. The separate threads correspond to respective text sources to enable the improved expansion query term generation to be provided in real time.

Brief Summary Text (13):
More specifically, the invention provides a dynamic statistical thesaurus including a collection of records which contain weighted term relationships. The statistical thesaurus is divided into multiple indexed collections based on sampled source material, and is searched interactively to construct a list of related concepts for one or more expanded query terms.

Brief Summary Text (15):
Moreover, the invention provides a statistical thesaurus structure which stores the collection of logical term relationship records as a document in which the terms are grouped by term weight in different indexed sections (segments) of the document. This allows a conventional document retrieval system to build the index, and create the candidate set of records for ranking.

Brief Summary Text (16):
The invention further provides a method of parallel processing which involves dividing the statistical thesaurus into small physical collections, each with its own index, searching the multiple collections simultaneously, and merging the search results.

Drawing Description Text (9):
FIG. 7 illustrates related concepts from a statistical (co-occurrence-based) thesaurus using news-based material.

Drawing Description Text (11):
FIG. 9 shows related concepts from a statistical thesaurus using GENFED (legal material) searches.

Drawing Description Text (12):
FIG. 10A illustrates an exemplary indexing scheme in a dictionary for a given collection, showing entries including a term in association with references to a document and a set of "groups" which reflect ranking of terms based on relevance.

Drawing Description Text (13):
FIG. 10B illustrates an exemplary "Top 100" List of ranked records, showing a "collection" number (based on text source), a document number, and a score (based on rankings determined by group within a record).

Drawing Description Text (15):
FIG. 11B illustrates a "Top 26" Terms Table showing terms ordered by frequency of occurrence.

Drawing Description Text (19):
FIG. 15 illustrates the relationship of the collections, text sources, threads, records (records correspond to documents), groups, and terms, according to a preferred embodiment of the statistical thesaurus according to the present invention.

Detailed Description Text (4):
For a statistical thesaurus, the related terms for a headword vary depending on the source text collection being searched, and over time as new material is added to the collection. Rebuilding a static list of related terms and headwords would be very computation-intensive and time consuming, limiting the ability to tune the thesaurus by source text collection and keep it current.

Detailed Description Text (7):
FIG. 1 illustrates a query expansion process. The process begins when the end user of the system enters a search query including one or more terms. The user may select to

expand the query, as a whole for statistical retrieval, or by specific term or terms for boolean retrieval. If the user specifies query expansion, the list of terms to be used for expansion is constructed, and the statistical thesaurus is accessed.

Detailed Description Text (10):
To summarize terminology, FIG. 15 illustrates the hierarchical relationship of the following terms according to a preferred embodiment of a statistical thesaurus according to the present invention:

Detailed Description Text (11):
The thesaurus includes plural collections, each collection being based on a respective test source (such as legal opinions, news stories, patent text, and so forth). The various collections are generated and searched in parallel, by respective (concurrently-executed) threads of a computer program. The collections include records. The records include groups of terms. The groups have weights (such as 1, 2, 3, 4 or 5) that constitute an indexing scheme that allows the user to interactively search the collections to generate query expansion terms.

Detailed Description Text (12):
The statistical thesaurus is a set of records, with each record containing a set of terms which are related to each other by their occurrence together in a body of text such as a document. The preferred embodiment of the invention designates five groups of terms in each record: Group 1 contains the most important terms from the body of text; group 2, the next most important terms; and so forth, through group 5 which contains the least important terms (although group 5 terms are still meaningful concepts within the body of text). These groupings in the document inherently reflect term weights for use in ranking the records during retrieval. The record may be generated by processing a body of text, and by extracting the important terms and phrases based on statistics using a suitable phrase recognition method such as that disclosed in application Ser. No. 08/589,468 which is incorporated herein by reference.

Detailed Description Text (14):
Significantly, the records are then grouped by the collections from which they were sampled. That way, the appropriate set of records can be accessed based on the collection selected by the user. For example, a first set of records may be formed based on federal case law documents, and a second set records may be formed based on news wires. When a user later searches case law, the first set of records is used for the statistical thesaurus. When the user is searching news material, the second set is used.

Detailed Description Text (16):
First, source documents are read, the valuable terms and phrases from the documents are extracted, and thesaurus "records" are written. The thesaurus records are essentially documents having a set of (for example) five groups (or document segments), each group inherently reflecting a ranking of the terms in the group.

Detailed Description Text (18):
GROUP1:

Detailed Description Text (20):
GROUP2:

Detailed Description Text (22):
GROUP3:

Detailed Description Text (23):
@ element of malice @ @ superfluous @ @ convicted of murder @ @ malice instruction @ @ degree murder @ @ habeas @ This simple example of a record contains terms only in the first three groups, due to the small size of the source document (the opinion in Davis v. State of Tennessee and Larry Lack, 856 F.2d 35; 1988 U.S. App. LEXIS 11941 (CA 6, 1988)). The "@" signs signal the beginning and end of terms to clearly delimit phrases. Of course, variations on this format lie within the contemplation of the present invention.

Detailed Description Text (24):
The thesaurus records are then processed to build a statistical thesaurus index and to build compressed records which are optimized for use in later retrieval operations. FIG. 10A illustrates an exemplary indexing scheme in a dictionary for a given collection, showing entries including a term in association with references to a

document and a set of "groups" which reflect ranking of terms based on relevance.

Detailed Description Text (26):
Furthermore, the index also specifies which term group the term is in. Each record can be thought of as a document. Each group can be though of as a sub-portion (or segment) of the document such as a paragraph. The records are grouped by their source collection type (legal or news), and exist in many different physical collections. A physical collection has its own index file and compressed text file.

Detailed Description Text (28):
FIG. 13 illustrates a preliminary process of selecting a collection which is to be searched in determining suitable query expansion terms. This process allows later processes to focus on a most appropriate source of phrases (case law, news wires, and so forth). First, the user-selected source is determined, and the source is looked up in a source map. A list of text source collections to be searched is then output. At this point, the processing illustrated in FIGS. 3 and 4 can take place.

Detailed Description Text (33):
The record is scored by tallying the "score" of the highest scoring location for each query term within the record, recognizing it is possible for a query term to appear multiple times within a record. Group 1 terms score 14 points, group 2 terms score 13 points, and so forth through group 5 terms which score 10 points. The maximum score for a record is 14 times the number of query terms. The minimum score is 10 times the number of query terms for boolean queries, and 10 for statistical queries.

Detailed Description Text (41):
The sort phase sorts the terms in the work file, and outputs the terms in alphabetical order. Multiple occurrences of the same term are now output consecutively. As they are output, the frequency of each term is calculated. A table of the top 26 terms, ranked by frequency, is maintained. FIG. 11B illustrates an exemplary "Top 26" Terms Table showing terms ordered by frequency of occurrence. In the preferred embodiment, a term must have a minimum frequency of 2 to be inserted into the table. After all sorted terms have been output, the Term Table is output to the end user.

Detailed Description Text (44):
In order to achieve maximum performance, several phases in the above process are performed in parallel. A parallel thread is created for each physical collection being searched. The statistical thesaurus is preferably built in several small collections instead of a single large collection. Preferably, these collections are based on the source of text, such as case law or news wires.

Detailed Description Text (47):
Thus, the invention provides a statistical thesaurus built from multiple information sources. Significantly, the statistical thesaurus is managed so that many different combinations of records may be searched to correspond to the collection specified by the end user.

Detailed Description Text (48):
As a particular example of the advantages of forming query expansion based on source text collection, a user of the LEXIS-NEXIS.TM. system may select the library GENFED (which contains federal case law), the library NEWS (which contains news media documents), or the library PATENT (which contains the full text of U.S. patents). These are examples of the source text collections mentioned above. If the user is searching in GENFED and the topic is MURDER, the related concepts provide better search performance if they are derived from federal case law. Conversely, the news media search would work better if the term is expanded using records generated from news documents. The difference in terms is clearly illustrated in FIGS. 7 and 9: FIG. 7 shows related concepts for NEWS searches, while FIG. 9 shows related concepts for GENFED searches.

Detailed Description Text (49):
This process is managed by sampling document collections individually, and then maintaining the generated term records in separate collections. Then, significantly, the term record collections are combined dynamically based upon the document collection being searched by the end user.

Detailed Description Text (50):
A hardware environment in which the inventive thesaurus may be developed, stored and used is shown in FIG. 14. In particular, a document search and retrieval system 30 is

shown. The system allows a user to search a subset of a plurality of documents for particular key words or phrases. The system then allows the user to view documents that match the search request. The system 30 comprises a plurality of Search and Retrieval (SR) computers 32-35 connected via a high speed interconnection 38 to a plurality of Session Administrator (SA) computers 42-44.

Detailed Description Text (57):
Each of the SA's 42-44 contains an application program that processes search requests input by a user at one of the terminals 64-66 and passes the search request information onto one or more of the SR's 32-35 which perform the search and returns the results, including the text of the documents, to the SA's 42-44. The SA's 42-44 provide the user with text documents corresponding to the search results via the terminals 64-66. For a particular user session (i.e. a single user accessing the system via one of the terminals 64-66), a single one of the SA's 42-44 will interact with a user through an appropriate one of the front end processors 56-58.

Detailed Description Text (58):
The collection selection method (FIG. 13) may be executed in either the session administrator SA computers 42-44 or in the search and retrieval computers 32-35. The remainder of the methods described above (FIGS. 3, 4) are preferably executed on the search and retrieval computers 32-35.

Detailed Description Paragraph Table (1):

		A.

Collections form the statistical thesaurus Text Sources are the basis of respective collections Threads in software can form and search respective collections B. Records in each collection are based on respective documents C. Groups of terms are found in each record D. Terms can include one or more words.

Other Reference Publication (3):
Peat, Helen J., et al., "The Limitations of Term Co-Occurrence Data for Query Expansion in Document Retrieval Systems", Journal of the American Society for Information Science, vol. 42, No. 5, 1991, pp. 378-383.

CLAIMS:

1. A system including a dynamic statistical thesaurus for use in interactively generating query expansion terms for use with an automated text document search and retrieval system, the system comprising:

a) means for receiving at least one search query term;

b) a plurality of collections of records, wherein:

b1) each record in a collection corresponds to a respective document;

b2) each record in a collection has term groups addressable by an indexing scheme;

b3) the collections are distinguished from each other based on respective text sample sources; and

b4) the term groups have different weights constituting part of the indexing scheme; and

c) means for using the indexing scheme to allow a user to interactively search the plurality of collections to generate the query expansion terms to supplement the at least one search query term.

2. The system of claim 1, wherein:

the search query term includes plural words constituting a phrase.

3. The system of claim 1, wherein:

the collections in the plurality of collections are searched concurrently using respective threads in parallel as subsets of a larger set of physically distributed collections.

11. A statistical thesaurus, comprising:

a plurality of collections of indexed records;

wherein:

1) the records constitute respective documents;

2) each document includes plural terms that are grouped by weight into different groups within the document; and

3) the groups are indexed so as to allow the records to be searched by a conventional text document search and retrieval system so as to perform functions of:

i) adding records to the indexed records and/or

ii) forming a list of related concepts for possible inclusion in expansion query terms.

12. The statistical thesaurus of claim 11, wherein

each document includes plural terms that are grouped by weight, from among about five possible different weights, into about five different respective groups within the document.

13. A query expansion method in a text document search and retrieval system using a statistical thesaurus to generate expansion query terms, the method comprising:

dividing the statistical thesaurus into multiple small physical collections, each physical collection having its own index;

searching the multiple collections in parallel, using the respective indexes; and

merging the search results to form a list of related concepts to be included in the expansion query terms.

19. The method of claim 13, wherein the searching step includes:

searching the multiple collections using threads in software.

Print Request Result(s)

Printer Name: cpk2_4c32_gbfrptr
Printer Location: cpk2__4c32

- US006208993: Ok
- US006175835: Ok
- US005983216: Ok
- US005926811: Ok
- US005794178: Ok
- US005619709: Ok

OK | Back to List | Logout

WEST

Freeform Search

Database: US Patents Full-Text Database
US Pre-Grant Publication Full-Text Database
JPO Abstracts Database
EPO Abstracts Database
Derwent World Patents Index
IBM Technical Disclosure Bulletins

Term: L15 and (occurrence near data)

Display: 50 **Documents in Display Format:** FRO **Starting with Number** 1

Generate: ○ **Hit List** ◉ **Hit Count** ○ **Side by Side** ○ **Image**

Search | Clear | Help | Logout | Interrupt

Main Menu | Show S Numbers | Edit S Numbers | Preferences | Cases

Search History

DATE: Saturday, October 19, 2002 Printable Copy Create Case

Set Name side by side	Query	Hit Count	Set Name result set
DB=USPT; PLUR=YES; OP=OR			
L16	L15 and (occurrence near data)	6	L16
L15	L14 and hierarch$	56	L15
L14	L13 and (search$ or organiz$)	124	L14
L13	L12 and (group$ or organiz$ or link$)	124	L13
L12	L11 and (search near term)	131	L12
L11	(search$ near retriev$) and occurrence	632	L11
DB=USPT,PGPB,JPAB,EPAB,DWPI,TDBD; PLUR=YES; OP=OR			
L10	L9 and (data near field$)	8	L10
L9	L8 and databas$	33	L9
L8	(occurrence near data) and (search$ near term$)	38	L8
L7	occurrence near data	2771	L7
L6	L4 and (occurrenc$)	0	L6
DB=USPT; PLUR=YES; OP=OR			
L5	L4 and (occurrence near data)	0	L5
L4	L2 or l1	2	L4
L3	L2 and l1	0	L3
L2	(6457006).pn.	1	L2
L1	(6424969).pn.	1	L1

END OF SEARCH HISTORY